数理解析研究所講究録
第1954巻 2015年 61-72

変分的手法に基づく尤度およびentropyの拡張

九州大学・経済学研究院　大西俊郎
Toshio Ohnishi
Faculty of Economics, Kyushu University

§1. Introduction

対数尤度とShannon entropyは統計学において最も基本的な量の2つである．モデルを$p_\theta(y)$とする．yが変数であり，θがパラメータである．データxが得られたとき，$\log p_\theta(x)$は対数尤度と呼ばれる．一方，確率密度$p(y)$に対し，Shannon entropyは

$$\mathrm{H}[p] := \mathrm{E}[-\log p(y) \,|\, p(y)]$$

によって定義される．ここで$\mathrm{E}[f(y) \,|\, p(y)]$は確率密度$p$の下での$f$の期待値を意味するものとする．

Amari & Nagaoka (2000)によれば，α-divergenceは凸関数

$$u^\alpha(t) := \begin{cases} -\log t & (\alpha = -1) \\ \frac{4}{1-\alpha^2}\left(1 - t^{\frac{1+\alpha}{2}}\right) & (-1 < \alpha < 1) \\ t\log t & (\alpha = 1) \end{cases}$$

を通じて

$$\mathrm{D}^\alpha(p,q) := \mathrm{E}\left[u^\alpha\left(\frac{q(y)}{p(y)}\right) \,\middle|\, p(y)\right]$$

のように定義される．ここで，u^αとD^αのαは上付きの添え字であって，べき乗ではない．

α-divergence は Kullback-Leibler divergence

$$\mathrm{KL}(p,q) = \mathrm{E}\left[\log \frac{p(y)}{q(y)} \,\middle|\, p(y)\right]$$

の一般化である．記号の約束として $\alpha = 1$ を e，$\alpha = -1$ を m と書く．つまり，

$$\mathrm{D}^e(p,q) = \mathrm{KL}(q,p), \quad \mathrm{D}^m(p,q) = \mathrm{KL}(p,q)$$

とする．本論文のねらいは，α-divergence を通じて対数尤度と Shannon entropy に統一的な視点を与えることである．

§2 以下で α について不連続なことが起きる場合があるが，これは u^α が α について不連続だからではない．実際，u^α を

$$\tilde{u}^\alpha(t) := \begin{cases} u^\alpha(t) + \frac{2}{1-\alpha}(t-1) & (-1 \le \alpha < 1) \\ u^\alpha(t) - (t-1) & (\alpha = 1) \end{cases}$$

のように修正しても凸性は変わらず，同一の α-divergence を表し，かつ，α に関して連続である．

§3 において等式が重要な役割を果たす．その等式は対数尤度または Shannon entropy と divergence の間に成立するものであり，Yanagimoto & Ohnishi (2009, 2011) は鞍点等式と名付けている．指数型分布族および混合分布を例題として取り上げ，どのような等式なのか説明しよう．

例題 1. θ を canonical パラメータとする次の指数型分布族を考える．

$$p_\theta(y) = \exp\{\theta y - \psi(\theta)\}a(y).$$

$\mu := \psi'(\theta)$ とおくと，これは平均パラメータである．x を任意に固定する．最尤推定量 MLE $\hat{\theta}$ を尤度方程式

$$\frac{\partial}{\partial \theta} \log p_\theta(x) = 0$$

の解によって定義すると，$x = \psi'(\hat{\theta})$ が得られる．次の等式が成立することが知られている (Kullback, 1959).

$$\log \frac{p_{\hat{\theta}}(x)}{p_\theta(x)} = \mathrm{D}^e\left(p_\theta, p_{\hat{\theta}}\right), \quad \forall (x,\theta).$$

実際，右辺と左辺を計算すると，ともに

$$\psi(\theta) - \psi(\hat{\theta}) - (\theta - \hat{\theta})x$$

となる．等式は対数尤度の差と e-divergence がバランスすることを意味する．

例題 2. p_1, p_2 は所与の確率密度とし，次の混合分布を考える．

$$p_\theta(y) = (1-\theta)p_1(y) + \theta p_2(y).$$

θ_* を方程式

$$\frac{\partial}{\partial\theta}\mathrm{H}\big[p_\theta\big] = 0$$

の解によって定義する．$(t\log t)' = 1 + \log t$ に注意して合成関数の微分を実行すると，θ_* を定義する方程式は

$$\int \{p_2(y) - p_1(y)\} \log p_{\theta_*}(y) dy = 0$$

と等価である．したがって，

$$\int \{p(y) - p_{\theta_*}(y)\} \log p_{\theta_*}(y) dy = 0$$

が成立する．Shannon entropy の定義を用いると，

$$\mathrm{H}\big[p_{\theta_*}\big] - \mathrm{H}\big[p_\theta\big] = \mathrm{D}^m\big(p_\theta, p_{\theta_*}\big), \quad \forall\theta$$

が得られる．この等式は，Shannon entropy の差と m-divergence のバランスを意味している．

2 つの例題は divergence と釣り合う量として対数尤度と Shannon entropy が自然に導かれることを暗示している．この暗示が正しいことが §3 で明らかにされる．

§2. 最小問題の定式化

本論文で対象とする最小問題は

$$\min_q \mathrm{E}\Big[\mathrm{D}^\alpha\big(p_\xi, q\big) \,\Big|\, h(\xi)\Big] \tag{2.1}$$

である．ここで，$p_\xi = p_\xi(y)$ は index ξ をもつ所与の確率密度であり，$h = h(\xi)$ は divergence の線形結合における重みを意味する確率密度である．この重み h を **canonical weight** と呼ぶことにする．

Bayes 予測問題が前頁の最小問題に帰着されることを示そう．Bayes モデル $p_\theta(y)\pi(\theta)$ において，データ x が得られたとき，予測分布 $q_x(y)$ によって真の分布 $p_\theta(y)$ を推定することを考える．損失関数を $\mathrm{D}^\alpha\big(p_\theta, q_x\big)$ とするとき，解くべき問題は

$$\min_q \mathrm{E}\Big[\mathrm{D}^\alpha\big(p_\theta, q_x\big) \;\Big|\; p_\theta(x)\pi(\theta)\Big]$$

である．大西 (2014) にあるとおり，この問題は

$$\min_q \mathrm{E}\Big[\mathrm{D}^\alpha\big(p_\theta, q_x\big) \;\Big|\; \pi_x(\theta)\Big] \tag{2.2}$$

と等価である．ただし，$\pi_x(\theta)$ は事後分布である．§1 で述べた本論文のねらいは，Ohnishi & Yanagimoto (2013) および大西 (2014) の理論のエッセンスを明らかにすることによって達成される．

本論文で重要な役割を果たす，確率密度の「算術平均」および「幾何平均」などを定義する．

Definition 1 (α-mixture). 確率密度 p_ξ の canonical weight h による平均を次式で定義し，**α-mixture** と呼ぶ．

- $-1 \le \alpha < 1$ のとき

$$f^\alpha[h](y) := \frac{1}{K^\alpha[h]}\Big[\mathrm{E}\big[\{p_\xi(y)\}^{\frac{1-\alpha}{2}} \mid h(\xi)\big]\Big]^{\frac{2}{1-\alpha}}$$

  特に，$\alpha = -1$ のとき $f^m[h](y)$ は「算術平均」である．

- $\alpha = 1$ のとき

$$f^e[h](y) := \frac{1}{K^e[h]}\exp\Big\{\mathrm{E}\big[\log p_\xi(y) \mid h(\xi)\big]\Big\}$$

  これは「幾何平均」である．

$f^\alpha[h](y)$ は h の汎関数であり，y の関数であるため，このような記号を用いている．規格化定数 $K^\alpha[h], K^e[h]$ も h の汎関数である．

以下，canonical weight の集合は Dirac のデルタ関数 $h(\xi') = \delta_D(\xi' - \xi)$ を含むと仮定する．定義から明らかなとおり，$h(\xi') = \delta_D(\xi' - \xi)$ のとき $f^\alpha[h] = p_\xi$ および $K^\alpha[h] = 1$ が成り立つ．

§3 で canonical weight h を微小に変化させた場合を考える．準備として次のように記号を定義しておく．具体的には，h_1, h_2 を 2 つの canonical weight とするとき h_1 を $h_1 + \beta(h_2 - h_1)$ のように変化させる．ここで β は $h_1 + \beta(h_2 - h_1)$ も確率密度になるような十分小さい数である．Canonical weight h の汎関数 $F[h]$ を考える．

Definition 2 (Gateaux 微分). h_1 における増分 $h_2 - h_1$ に対する汎関数 $F[h]$ の **Gateaux 微分**を

$$\delta_G F[h_1; h_2 - h_1] := \lim_{\beta \to 0} \frac{F\big[h_1 + \beta(h_2 - h_1)\big] - F[h_1]}{\beta}$$

によって定義する．

§3. 4 つの定理

本論文の主張のエッセンスは以下の 4 つの定理に凝縮される．

Theorem 1 (最適解). Definition 1 の α-mixture $f^\alpha[h]$ は §2 の最小問題 (2.1) の最適解である．

証明：ここでは $-1 \leq \alpha < 1$ の場合を証明する．α-mixture の定義から得られる等式

$$\big\{K^\alpha[h]\, f^\alpha[h](y)\big\}^{\frac{1-\alpha}{2}} = \mathrm{E}\Big[\big\{p_\xi(y)\big\}^{\frac{1-\alpha}{2}} \,\Big|\, h(\xi)\Big] \tag{3.1}$$

に注意する．α-divergence の差の期待値を計算すると，

$$\mathrm{E}\Big[\mathrm{D}^\alpha\big(p_\xi, q\big) - \mathrm{D}^\alpha\big(p_\xi, f^\alpha[h]\big) \,\Big|\, h(\xi)\Big] = \big\{K^\alpha[h]\big\}^{\frac{1-\alpha}{2}}\, \mathrm{D}^\alpha\big(f^\alpha[h], q\big)$$

が得られる．右辺は明らかに非負である．□

Theorem 2 (平均鞍点等式). 最適解は次の等式を満たす．

- $\alpha = -1$ のとき

$$\mathrm{E}\Big[\mathrm{H}\big[f^m[h]\big] - \mathrm{H}\big[p_\xi\big] - \mathrm{D}^m\big(p_\xi,\, f^m[h]\big) \;\Big|\; h(\xi)\Big] = 0.$$

- $-1 < \alpha \leq 1$ のとき

$$\mathrm{E}\left[u^{-\alpha}\left(\frac{p_\xi(x)}{f^\alpha[h](x)}\right) - \mathrm{D}^\alpha\big(p_\xi, f^\alpha[h]\big) \;\middle|\; h(\xi)\right] = 0, \quad \forall x.$$

$-1 < \alpha \leq 1$ のときのみ「データ x」が現れ，$\alpha = -1$ のときは現れないことに注意されたい．

証明：ここでは $-1 < \alpha < 1$ のときのみ証明する．Theorem 1 の証明と本質的に同じであり，(3.1) から導ける．□

Defintion 3 (Divergence 共役量). x を任意に固定する．Theorem 2 の等式において divergence 損失と平均的にバランスしている量を **divergence 共役量**と呼ぶ．

- $\alpha = -1$ のときは Shannon entropy の差

$$\mathrm{H}\big[f^m[h]\big] - \mathrm{H}\big[p_\xi\big]$$

  である．この場合は x に依存しない．

- $-1 < \alpha \leq 1$ のときは尤度比を関数 $u^{-\alpha}$ で変換したものである．

$$u^{-\alpha}\left(\frac{p_\xi(x)}{f^\alpha[h](x)}\right).$$

  特に $\alpha = 1$ のとき，対数尤度比 $\log\{f^\alpha[h](x)/p_\xi(x)\}$ となる．

$-1 < \alpha \leq 1$ のとき尤度比を変換する関数が u^α でないことに注意されたい．

Theorem 2 は，canonical weight の下での divergence 共役量の期待値が最小問題 (2.1) の最小値と一致することを意味する．(3.1) から，

$$\mathrm{E}\left[u^{-\alpha}\left(\frac{p_\xi(x)}{f^\alpha[h](x)}\right)\middle|\, h(\xi)\right]$$

は $u^{-\alpha}\big(K^{\alpha}[h]\big)$ に等しく，x に依存しないことに注意する．

Defintion 4 (最小問題の最小値). 最小問題 (2.1) の最小値を $-\psi^{\alpha}[h]$ とおく．具体的には，

- $\alpha=-1$ のとき
$$-\psi^{m}[h] := \mathrm{H}\big[f^{m}[h]\big] - \mathrm{E}\Big[\mathrm{H}\big[p_{\xi}\big] \;\Big|\; h(\xi)\Big].$$
- $-1<\alpha\leq 1$ のとき
$$-\psi^{\alpha}[h] := u^{-\alpha}\big(K^{\alpha}[h]\big).$$

Theorem 3 (Divergence 共役量の Gateaux 微分). Definition 3 の divergence 共役量の Gateaux 微分は次のとおり．

- $\alpha=-1$ のとき
$$-\Big\{\mathrm{H}\big[f^{m}[h_1]\big] - \mathrm{H}\big[f^{m}[h_2]\big] - \mathrm{D}^{m}\big(f^{m}[h_2], f^{m}[h_1]\big)\Big\}.$$
- $-1<\alpha\leq 1$ のとき
$$-\left\{\frac{K^{\alpha}[h_2]}{K^{\alpha}[h_1]}\frac{p_{\xi}(x)}{f^{\alpha}[h_1](x)}\right\}^{\frac{1-\alpha}{2}}\left\{u^{-\alpha}\left(\frac{f^{\alpha}[h_2](x)}{f^{\alpha}[h_1](x)}\right) - \mathrm{D}^{\alpha}\big(f^{\alpha}[h_2], f^{\alpha}[h_1]\big)\right\}.$$

証明：ここでは $\alpha=-1$ のときのみ証明を与える．$f^{m}[h]$ の Gateaux 微分を計算すると，

$$\delta_{G} f^{m}[h_1; h_2-h_1] = f^{m}[h_2] - f^{m}[h_1]$$

となる．$(t\log t)' = 1+\log t$ に注意して合成関数の微分を実行すると，証明すべき結果が得られる．□

Canonical weight の集合の中に divergence 共役量を停留させる canonical weight が存在すると仮定する．具体的には次のような canonical weight が存在すると仮定する．

- $\alpha=-1$ のとき
$$h^{m\dagger} := \operatorname*{argext}_{h}\big\{\mathrm{H}\big[f^{m}[h]\big] - \mathrm{H}\big[p_{\xi}\big]\big\}.$$

- $-1 < \alpha \leq 1$ のとき

$$h_x^{\alpha\dagger} := \operatorname*{argext}_h u^{-\alpha}\left(\frac{p_\xi(x)}{f^\alpha[h](x)}\right).$$

Theorem 3 において $h_2(\xi') = \delta_D(\xi' - \xi)$ とすると次の系を得る.

Corollary to Theorem 3 (Exact な鞍点等式). 上のような canonical weight が存在するとき, 次の等式が成り立つ.

- $\alpha = -1$ のとき

$$\mathrm{H}\big[f^m[h^{m\dagger}]\big] - \mathrm{H}\big[p_\xi\big] = \mathrm{D}^m\big(p_\xi,\, f^m[h^{m\dagger}]\big).$$

- $-1 < \alpha \leq 1$ のとき

$$u^{-\alpha}\left(\frac{p_\xi(x)}{f^\alpha[h_x^{\alpha\dagger}](x)}\right) = \mathrm{D}^\alpha\big(p_\xi,\, f^\alpha[h_x^{\alpha\dagger}]\big),\ \ \forall x.$$

これらは divergence 共役量を停留させると divergence とその共役量が一致することを意味し, §1 の例題 1 および 2 の一般化になっている. Theorem 2 および 3 から, 尤度と Shannon entropy を事前分布 $\pi(\theta)$ の汎関数としてとらえることでこれらの概念の拡張が可能であることが分かる. Bayes モデル $p_\theta(y)\pi(\theta)$ およびデータ x が与えられたとき, 普通は最小問題 (2.2) を考えるが, 敢えてここでは

$$\min_q \mathrm{E}\Big[\mathrm{D}^\alpha\big(p_\theta, q\big) \;\Big|\; \pi(\theta)\Big]$$

を考察する. Theorem 2 および 3 から, 尤度の拡張は

$$f^e[\pi](x) = \frac{1}{K^e[\pi]} \exp\big\{\mathrm{E}[\log p_\theta(x)|\pi(\theta)]\big\}$$

によって可能であり, Shannon entropy の拡張は

$$\mathrm{H}\big[f^m[\pi]\big] = \mathrm{H}\big[\mathrm{E}[p_\theta|\pi(\theta)]\big]$$

によって可能であることが分かる. この両者は次の確率密度の「平均」である α-mixture を通じて統一的に理解できる.

$$f^\alpha[\pi](x) = \frac{1}{K^\alpha[\pi]}\Big[\mathrm{E}\big[\{p_\theta(x)\}^{\frac{1-\alpha}{2}} \;\big|\; \pi(\theta)\big]\Big]^{\frac{2}{1-\alpha}}.$$

Theorem 3 とほぼ同様の計算により，次の定理が得られる．

Theorem 4 (最小問題の最小値の Gateaux 微分). Definition 4 の最小値 $-\psi^{\alpha}[h]$ の Gateaux 微分は次のようになる．

$$-\delta_G\psi^{\alpha}[h_1; h_2 - h_1] = \mathrm{E}\Big[\mathrm{D}^{\alpha}\big(p_{\xi}, f^{\alpha}[h_1]\big) \;\Big|\; h_2(\xi) - h_1(\xi)\Big].$$

最小問題の最小値 $-\psi^{\alpha}_x[h]$ を停留させる canonical weight が存在すると仮定する．

$$h^{\alpha\mathrm{c}} := \operatorname*{argext}_{h}\big\{-\psi^{\alpha}_x[h]\big\}.$$

Theorem 4 において $h_2(\xi') = \delta_D(\xi' - \xi)$ とすると次の系を得る．

Corollary to Theorem 4 (リスク一定). 任意の canonical weight $h(\xi)$ に対して

$$\mathrm{E}\Big[\mathrm{D}^{\alpha}\big(p_{\xi},\, f^{\alpha}[h^{\alpha\mathrm{c}}]\big) \;\Big|\; h(\xi)\Big]$$

は一定である．

§4. 熱力学原理とのアナロジー

発見的ではあるが，canonical weight と対になる量を定義する．

Definition 5 (Mean weight). x を任意に固定する．Canonical weight h の汎関数であり，かつ，ξ の関数である次の量を **mean weight** と呼ぶ．

- $-1 \leq \alpha < 1$ のとき

$$t^{\alpha}_x[h](\xi) := u^{\alpha}\big(f^{\alpha}[h](x)\big) - \mathrm{D}^{\alpha}\big(p_{\xi}, f^{\alpha}[h]\big).$$

- $\alpha = 1$ のとき

$$t^{e}[h](\xi) := -\mathrm{H}\big[f^{e}[h]\big] - \mathrm{D}^{e}\big(p_{\xi},\, f^{e}[h]\big).$$

Mean weight と canonical weight の関係は，指数型分布族における平均パラメータと正準パラメータの関係と似ている．実際に，Definition 4 で定義した最小値 $-\psi^\alpha[h]$ を用いると次の関係式が成り立つ．

- $-1 \leq \alpha < 1$ のとき

$$\delta_G \psi^\alpha[h_1; h_2 - h_1] = \mathrm{E}\big[t_x^\alpha[h_1](\xi) \;\big|\; h_2(\xi) - h_1(\xi)\big].$$

- $\alpha = 1$ のとき

$$\delta_G \psi^e[h_1; h_2 - h_1] = \mathrm{E}\big[t^e[h_1](\xi) \;\big|\; h_2(\xi) - h_1(\xi)\big].$$

また，指数型分布族におけるキュムラント関数が凸関数であるのと同様に汎関数 $\psi^\alpha[h]$ の凸性を証明することができる．

Theorem 5 (最小問題の最小値の凹性). Definition 4 で定義された $\psi^\alpha[h]$ は凸汎関数である．

証明：$0 \leq \beta \leq 1$ とする．$g := (1-\beta)h_1 + \beta h_2$ とおき，canonical weight g に対する最適解を $f^\alpha[g]$ とおく．$-\psi^\alpha[g] = \mathrm{E}\big[\mathrm{D}^\alpha\big(p_\xi, f^\alpha[g]\big) \;\big|\; g\big]$ は次のように不等式で評価できる．

$$\begin{aligned}\mathrm{E}\big[\mathrm{D}^\alpha\big(p_\xi, f^\alpha[g]\big) \;\big|\; g\big] &= (1-\beta)\mathrm{E}\big[\mathrm{D}^\alpha\big(p_\xi, f^\alpha[g]\big) \;\big|\; h_1\big] + \beta\mathrm{E}\big[\mathrm{D}^\alpha\big(p_\xi, f^\alpha[g]\big) \;\big|\; h_2\big] \\ &\geq (1-\beta)\mathrm{E}\big[\mathrm{D}^\alpha\big(p_\xi, f^\alpha[h_1]\big) \;\big|\; h_1\big] + \beta\mathrm{E}\big[\mathrm{D}^\alpha\big(p_\xi, f^\alpha[h_2]\big) \;\big|\; h_2\big].\end{aligned}$$

この不等式は $-\psi^\alpha[h]$ の凹性を示している．□

最小問題 (2.1) は制約条件なしの最小問題である．これを，同一の最適解 $f^\alpha[h]$ をもつ制約条件つき最大問題に書き換える．

Theorem 6 (同一の最適解をもつ制約つき最大問題).

- $-1 \leq \alpha < 1$ のとき

  最小問題 (2.1) と次の問題は，$s_x(\xi) = t_x^\alpha[h](\xi)$ のときに限り，同一の最適解 $f^\alpha[h]$

をもつ.

$$\begin{aligned} &\max \ -u_\alpha\big(q(x)\big) \\ &\text{subject to } \ u_\alpha\big(q(x)\big) - \mathrm{D}^\alpha\big(p_\xi, q\big) = s_x(\xi) \end{aligned}$$

- $\alpha = 1$ のとき

  最小問題 (2.1) と次の問題は，$s(\xi) = t^e[h](\xi)$ のときに限り，同一の最適解 $f^\alpha[h]$ をもつ.

$$\begin{aligned} &\max \mathrm{H}\big[q\big] \\ &\text{subject to } \ -\mathrm{H}\big[q\big] - \mathrm{D}^e\big(p_\xi, q\big) = s(\xi) \end{aligned}$$

証明の本質：$d(A,B)$ を 2 つの点 A, B の乖離度とするとき，$d(A,X)$ と $d(B,X)$ を同時に小さくすることを考える．適当に h を決め，$(1-h)d(A,X) + hd(B,X)$ を最小化する．$d(B,X) - d(A,X) = t$ を固定し，$d(A,X)$ を最小化する．両者は Lagrange の未定乗数法で結ばれている．□

Theorem 5 は，状況に応じて原理を等価変形する熱力学に似ている．熱力学において次の 2 つの原理

- Energy minimum principle:

  Entropy が一定のとき，平衡状態では内部エネルギーが最小化される.

- Helmholtz potential minimum principle:

  温度が一定のとき，平衡状態では Helmholtz potential が最小化される.

は等価であり，内部エネルギーと Helmholtz potential が Legendre 変換で結ばれていることが知られている.

REFERENCES

Amari, S-I. and Nagaoka, H. (2000). *Methods of Information Geometry.* American

Mathematical Society, Load Island.

Kullback, S. (1959). *Information Theory and Statistics.* Wiley, New York.

Ohnishi, T. and Yanagimoto, T. (2013). Twofold structure of duality in Bayesian model averaging. *Journal of the Japan Statistical Society, to appear.*

Yanagimoto, T. and Ohnishi, T., (2009). Bayesian prediction of a density function in terms of e-mixture. *Journal of Statistical Planning and Inference*, **139**, 3064-3075.

Yanagimoto, T. and Ohnishi, T., (2011). Saddlepoint condition on a predictor to re-confirm the need for the assumption of a prior distribution. *Journal of Statistical Planning and Inference*, **141**, 1990-2000.

大西俊郎, (2014). Bayes 予測における尤度とエントロピーの双対性. 京都大学 数理解析研究所 講究録, **1910**, 29–42.